Merging Segmental And Rhythmic Features For Automatic Language Identification

Jérôme Farinas, François Pellegrino, Jean-Luc Rouas, Régine André-Obrecht

▶ **To cite this version:**

Jérôme Farinas, François Pellegrino, Jean-Luc Rouas, Régine André-Obrecht. Merging Segmental And Rhythmic Features For Automatic Language Identification. International Conference on Acoutics, Speech and Signal Processing, 2002, Orlando, Florida, United States. pp.753-756. hal-00702443

HAL Id: hal-00702443
https://hal.science/hal-00702443

Submitted on 5 Jun 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MERGING SEGMENTAL AND RHYTHMIC FEATURES FOR AUTOMATIC LANGUAGE IDENTIFICATION

Jérôme FARINAS[1], *François PELLEGRINO*[2], *Jean-Luc ROUAS*[1] *and Régine ANDRE-OBRECHT*[1]

[1]Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS INPT UPS, FRANCE
[2]Laboratoire Dynamique Du Langage, UMR 5596 CNRS Univ. Lyon 2, FRANCE

ABSTRACT

This paper deals with an approach to Automatic Language Identification based on rhythmic modeling and vowel system modeling. Experiments are performed on read speech for 5 European languages. They show that rhythm and stress may be automatically extracted and are relevant in language identification: using cross-validation, 78% of correct identification is reached with 21 s. utterances. The Vowel System Modeling, tested in the same conditions (cross-validation), is efficient and results in a 70% of correct identification for the 21 s. utterances. Last, merging the two models slightly improves the results.

1. INTRODUCTION

During the last decade, the request for Automatic Language Identification (ALI) systems has arisen in several fields of application, and especially in Computer-Assisted Communication (e.g. Emergency Service) and Multilingual Man-Computer Interfaces (e.g. Interactive Information Terminal). More recently, content-based indexing of multimedia or audio data has provided a new topic in which ALI systems may be useful. However, current ALI systems are still not efficient enough to be used in a commercial framework. In the standard up to date approach, sequences of phonetic units (provided by a phonetic modeling system) are decoded according to language-specific statistical grammars [1]. This approach, initiated at the beginning of the 90s, is still the most efficient one. However, only marginal improvements have been performed for five years, and it seems crucial to propose new approaches. In this paper, we investigate the way to explicitly take phonetics into account and to take advantage from alternative features also present in the signal: prosodic features, and especially rhythmic features, are known to carry a substantial part of the language identity (Section 2). However, their modeling is still an open problem, mostly because of the nature of the prosodic features. To address this problem, an algorithm of language independent extraction of rhythmic features is proposed and applied to model rhythm (Section 3). This algorithm, coupled with a Vowel System Model (VSM) is tested on the five languages of the MULTEXT corpus in section 4. The relevance of the rhythmic parameters and the efficiency of each system (Rhythmic Model and Vowel System Model) are evaluated. Furthermore, the possibility of merging these two approaches is addressed.

2. MOTIVATIONS

2.1. Relevancy of Rhythm

Rhythm is a characteristic of language that is critical in different activities related to language (e.g. child language acquisition, language synthesis), and especially in both human and computer language identification. Among others, Thymé-Gobbel and Hutchings point out the importance of prosodic information in language identification systems [2]. With parameters related to rhythm and based on syllable timing, syllable duration, and descriptors of amplitude patterns, they have obtained promising results, and proved that mere prosodic cues can distinguish between some language pair with results comparable to some non-prosodic systems. Ramus et al. [3] show that newborn infants are sensitive to the rhythmic properties of languages. Other experiments based on a consonant/vowel segmentation of eight languages established that derived parameters might be relevant to classify languages according to their rhythmic properties [4].

2.2. Classifying languages according to rhythm

Experiments reported here focus on 5 European languages (English, French, German, Spanish and Italian). According to the literature, French, Spanish and Italian are “syllable-timed” while English and German are “stress-timed”. These two categories emerged from the theory of isochrony introduced by Pike and developed by Abercrombie [5]. However, more recent works based on the measurement of the duration of inter-stress intervals in both stress-timed and syllable-timed languages provide an alternative framework in which these two binary categories are replaced by a continuum [6]. Rhythmic differences between languages are

then mostly related to their syllable structure and the presence (or absence) of vowel reduction. The controversies on the status of rhythm in world languages illustrate dramatically the difficulty to segment speech into correct rhythmic units. Even if correlates between speech signal and linguistic rhythm exist, reaching a relevant representation seems difficult. Another difficulty rises from the selection of an efficient modeling paradigm. We develop here a statistical approach, first introduced in [7] and now improved by considering stress features (F_0 and Energy). It is based on a Gaussian modeling of the different "rhythm units" automatically extracted from a rhythmic segmentation in the languages.

3. DESCRIPTION OF THE SYSTEM

A synopsis is displayed in Figure 1. A language independent vowel detection algorithm is applied to label the speech signal in Silence/Non Vowel/Vowel segments. Afterward, computation of cepstral coefficients for the vowel segments leads to language-specific Vowel System Models (VSM), while the rhythmic pattern derived from the segmentation is used to model the rhythm of each language.

3.1. The Vowel/Non Vowel segmentation algorithm

This algorithm, based on a spectral analysis of the signal, is described in [8]. It is applied in a language and speaker independent way without any manual adaptation phase. This processing provides a segmentation of the speech signal in pause, non-vowel and vowel segments. Due to the intrinsic properties of the algorithm (and especially the fact that transient and steady parts of a phoneme may be separated), it is somewhat incorrect to consider that this segmentation is exactly a Consonant/Vowel segmentation. However, it is undoubtedly correlated to the rhythmic structure of the speech sound, and in this paper, we investigate the assumption that this correlation enables a statistical model to discriminate languages according to their rhythmic structure.

3.2. Vowel System Modeling

Each vowel segment is represented with a set of 8 Mel-Frequency Cepstral Coefficients and 8 delta-MFCC, augmented with the Energy and delta Energy of the segment. This parameter vector is extended with the duration of the underlying segment providing a 19-coefficient vector.

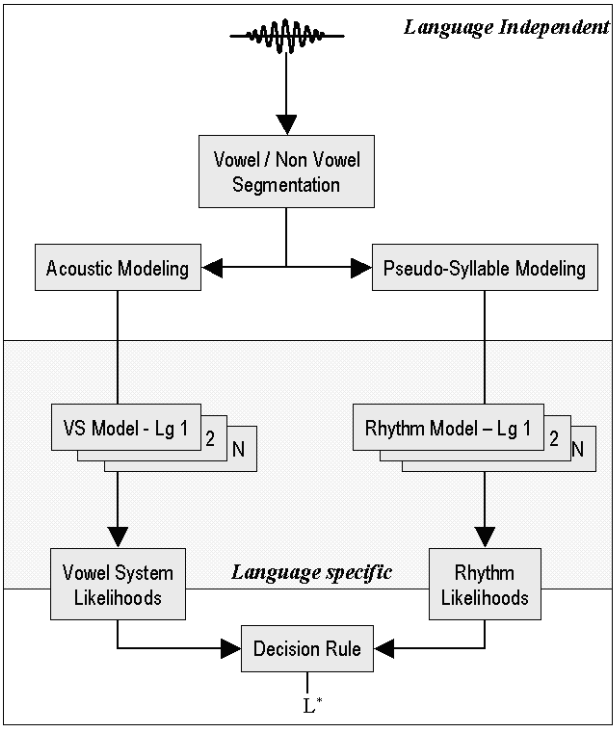

Figure 1 - Synopsis of the system for N languages.

A cepstral subtraction performs both blind removal of the channel effect and speaker normalization. For each recording sentence, the average MFCC vector is computed and subtracted from each coefficient. For each language, a Gaussian Mixture Model (GMM) is trained using the EM algorithm. The number of components of the model is computed using the LBG-Rissanen algorithm [9]. During the test, the decision lays on a Maximum Likelihood procedure.

3.3. Rhythm Modeling

3.3.1. Rhythmic units

Syllable may be a first-rate candidate for rhythm modeling. Unfortunately, segmenting speech in syllables is typically a language-specific mechanism and thus no language independent algorithm can be derived. For this reason, we introduced in [7] the notion of pseudo-syllables derived from the most frequent syllable structure in the world, namely the CV structure [10]. In the algorithm, speech signal is parsed in patterns matching the structure: .CnV. (where n is an integer that may be zero and V may result from the merging of consecutive vowel segments). For example, if the vowel detection algorithm results in the sequence (CCVVCCVCVC-CCVCVCCC), it is parsed in the following sequence of 5 pseudo-syllables: (CCV.CCV.CV.CCCV.CV)

3.3.2. Pseudo-syllable description

For each pseudo-syllable, three parameters are computed, corresponding respectively to the total consonant cluster duration, the total vowel duration and the complexity of the consonantal cluster. For example, the description for a .CCV. pseudo-sequence is:

$$P_{.CCV.} = \{D_C \; D_V \; N_C\}$$

where D_C is the total duration of the consonantal segments, D_V is the duration of the vowel segment and N_C is the number of segments in the consonantal cluster (here, N_C = 2). Additionally, two parameters related to the stress structure of the language (F_0 and Energy in dB, normalized among the sentence) are also considered. Our hypothesis is that these parameters may improve the discrimination of stress-timed languages. Such a basic rhythmic parsing is obviously limited, but it provides a framework to model rhythm that requires no knowledge on the language rhythmic structure

3.3.3. Statistical Rhythm modeling

For each language, a GMM is trained, either by using the standard LBG algorithm or the LBG-Rissanen algorithm to provide the optimal number of Gaussian components.

4. EXPERIMENTS

4.1. Corpus

Experiments are performed on the MULTEXT corpus [1]. This database contains recordings from five European languages (English, French, German, Italian and Spanish), pronounced by 50 different speakers (5 male and 5 female per language). Data consist of read passages of about five sentences extracted from the EUROM1 speech corpus (the mean duration of each passage is 20.8 seconds). The raw pitch contour of the signal is also available. A limitation is that the same texts are produced on average by 3.75 speakers, resulting in a possible partial text dependency of the models. Due to the limited size of the corpus, language identification experiments are performed using a cross-validation procedure: 9 speakers are used for training the models of one language and the tenth speaker is used for test. This procedure is iterated for each speaker, and for each language.

4.2. Rhythm Modeling

Table 1 summarizes the experiments performed with the rhythm parameters. The identification scores displayed are averaged among several GMM topologies and obtained using the whole duration of the test excerpts (about 21 seconds).

Table 1 - Results in cross-validation experiments with rhythm modeling.

Parameters	Mean Identification Rate
D_V + D_C	64.8 %
D_V + D_C + N_C	70.0 %
D_V + D_C + N_C + E	75.0 %
D_V + D_C + N_C+ E + F_0	69.4 %

The use of duration parameters D_V and D_C results in a 64.8 % of correct identification. The use of additional parameters related to the complexity of the pseudo-syllable structure (N_C) and to the stress (E) significantly improves the results, reaching 75 % of correct identification. In contrast, F_0 does not improve the results. This result may signify that a static value of F_0 per pseudo-syllable, even if it is normalized, is not significant enough to be useful. In another experiment (see Figure 2), influence of duration of test excerpts is tested. Modeling is performed in the four dimension space (D_V + D_C + N_C + E) which is the most efficient.

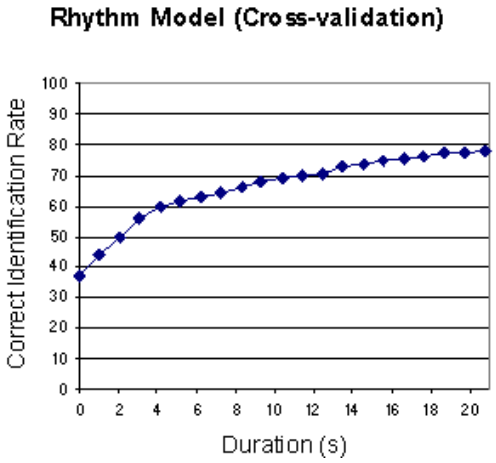

Figure 2 - Correct Identification rate in function of duration of test excerpts (Rhythm Model).

Unsurprisingly, identification rate increases with test excerpt duration to reach about 78 % with 21 s. However, even with short test utterances (less than ten pseudo-syllables), results are much more than chance. Furthermore, using only the first pseudo-syllable of the sentence results in a 37 % of correct identification (to be compared to chance: 20 %).

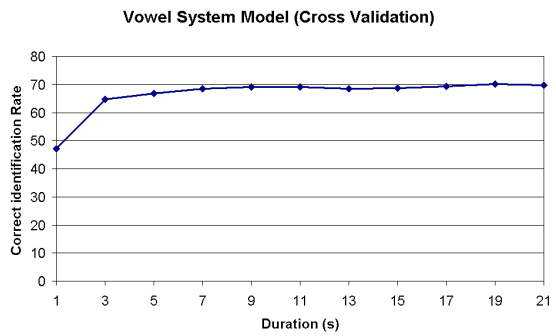

Figure 3 - Correct Identification rate in function of duration of test excerpts (Vowel System Model).

4.3. Vowel System Modeling

As shown in Figure 3, the Vowel system modeling approach is efficient with the MULTEXT corpus. An identification level of 42 % is reached with 1 second of signal. Increasing duration of test utterances allows reaching 70 % of correct identification for 21 seconds.

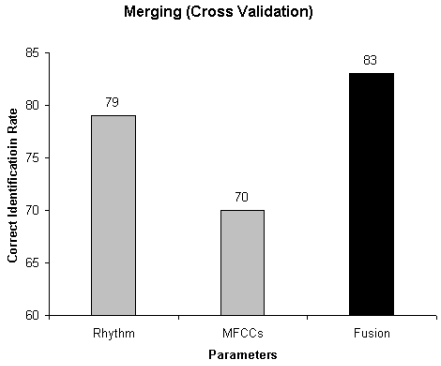

Figure 4 - Best Correct Identification rate for VSM, Rhythm model, and merging of the approaches.

4.4. Integrating Rhythm and Segmental Modeling

A simple statistical merging is performed by adding the log-likelihoods of both the Rhythm model and the VSM for each language. The scores obtained with 21seconds utterances are displayed in Figure 4. Merging the two approaches allows to reach 83 % of correct identification.

5. DISCUSSION

We propose in this paper two algorithms dedicated to automatic language identification. Experiments, performed with cross-validation, show that it is possible to achieve an efficient rhythmic modeling (78% of correct identification) in a way that requires no a priori knowledge of the rhythmic structure of the processed languages. Besides, the Vowel System Model reaches 70 % of correct identification. With these read data, merging the two approaches improves the identification rate up to 83 %.

6. ACKNOWLEDGEMENTS

This research is supported by the EMERGENCE program of the Région Rhône-Alpes and the French *Ministère de la Recherche* (program ACI *Jeunes Chercheurs*).

7. REFERENCES

[1] Zissman, M. A., Berkling, K. M., *Automatic language identification*, Speech Communication, Vol. 35, no. 1-2, pp. 115-124, 2001.

[2] Thymé-Gobbel, A., and Hutchins, S. E., *Prosodic features in automatic language identification reflect language typology*, Proc. of ICPhS99, San Francisco, 1999.

[3] Ramus, F., Hauser, M. D., Miller, C., Morris, D. and Mehler, J., *Language discrimination by human newborns and by cotton-top tamarin monkeys*, Science, 288, 349-351, 2000.

[4] Ramus, F., Nespor, M., & Mehler, J., *Correlates of linguistic rhythm in the speech signal*, Cognition, 73(3), 265-292, 1999.

[5] Abercrombie, D., Elements of General Phonetics, Edinburgh University Press, Edinburgh, 1967.

[6] Dauer, R. M., *Stress-timing and syllable-timing reanalyzed*, Journal of Phonetics, 11:51-62, 1983.

[7] Farinas, J. and Pellegrino, F., *Automatic Rhythm Modeling for Language Identification*, Proc. Of Eurospeech Scandinavia 01, Aalborg, 2001.

[8] Pellegrino, F., and André-Obrecht, R., *An Unsupervised Approach to Language Identification*, Proc. of ICASSP99, Phoenix, 1999.

[9] Pellegrino, F. and André-Obrecht, R., *Automatic Language Identification: an Alternative Approach to Phonetic Modeling*, Signal Processing, 80, 2000.

[10] Vallée, N., Boë, L.J., Maddieson, I. and Rousset, I., *Des lexiques aux syllabes des langues du monde : Typologies et structures*, Proc. of JEP 2000, Aussois, 2000.

[11] Campione, E., and Véronis, J., *A multilingual prosodic database*, Proc. of ICSLP'98, Sidney, 1998.